**Report**

**Team member:**

**颜渤东 119010369 毛邦靖 118020314 杨洪萌 119010378**

**Choice of Dataset: Dataset2: Quality of red wine**

1. **The data and the question that we are interested in answering.**

We choose the classification of dataset 1(Google Play Store Apps) at the beginning, and we write a super big code, but we failed. **Finally, we try to do the classification of dataset 2 (Red Wine Quality), and we got the answer. This report will show the detail of the code of the final version.**

1. **The review of some of the approaches that we tried or thought about trying**

We learned three methods ID3, C4.5 and CART, but finally we chose CART. In the following part, we will show how we explore these three methods and make decision.

First, when we first learn decision tree, we found three models ID3, C4.5 and CART. In order to make it easier, we choose to use the method from the easier one to harder one.

Second, we chose ID3 first, the main idea of ID3 is to use information gain to choose internal node, but when we tried to use this method to solve this problem, we found the data of the properties of red wine is continuous type. However, the ID3 can only solve discontinuous type data. When we try to use it to solve some discontinuous type data, we found the accuracy of this method is not very high, because we found the information gain cannot stand for the degree of reduction of uncertainty completely. We found that if the more the value of a property, the bigger the information gain is. And it cannot use to solve data has unknown value, so we give up this method.

Third, we tried C4.5 second, in order to solve the problem in ID3, C4.5 has changed some important part. For example, C4.5 use the rate of information gain in order to solve problem of the more the value of a property, the bigger the information gain is. However, we also found some drawbacks. First, using C4.5 maybe create a tree but not a binary tree, which may decrease the efficiency of the program, and because C4.5 uses the model of entropy there are a lot of calculation of the logarithmic operation which also will decrease the accuracy and efficiency. Therefore, we chose use CART at last.

And the first time we tried to do the classification of dataset 1 because it the most difficult question, and we want to use what we have learn to solve it and we write a super big file, but we failed. **(if you want to see the super code, don’t hesitant e-mailing us, we also have the report of that code)** Then we change our direction to do the classification of dataset 2, and we got the final answer.

1. **The summary of the final approach we used and why we chose this approach.**

At last, our group chose the CART method

The reason why we choose this method:

1. CART use Gini index which can decrease the calculation of the whole program, and the calculation of Gini index have no logarithmic operation, so it may be more accurate.
2. Gini index can solve the problem od both continuous and discontinuous type of data
3. The method of cut the tree is post-cutting method, which can choose the best cutting way in order to get the best generalization ability
4. Because the calculation of the Gini index is easier than the calculation of entropy, so the efficiency of the program will be better. And it can be faster.

Total method of our method:

1. Processing data: Because the data of red wine has many types, in order to solve problem easier, we want to process data first.
2. Create tree: We use the function we define by ourselves to calculate Gini index first. After that we according to the Gini we have to decide the root and internal node step by step.
3. Cut tree: After we get the tree, we will cut the tree in order to get the most efficient way to get the result, we need to simplify the tree by using cut some part of tree which can increase the performance of the program. We do precutting in this program
4. Predict the result: After we get the classification tree, we need to use the test file to test the performance of the program, at the same time, calculate the accuracy.
5. **The explanation the working principle and logic of the approach used**

**datasplit():**This function is used to read the train file and sort the data. Every data and its information are a list, and put them into a list named settt.

We use split function to let the information into several parts. Then return the list settt which is a matrix of the data set of train file.

**continuous\_gini():**This function is used to calculate the Gini of continuous type data.

ginis is a list which record the Gini value of the split point in order to do comparation of the Gini.

The program split the certain into 4 pieces, the two conditions are whether this value greater or equal to the split point value, and the equality is greater or equal to 6 or not. And using these four parts to calculate the Gini value of this split point.

After get the Gini value of split point the program will do comparation of the Gini value and the smallest Gini value we have now. If this Gini value is smaller, then the program will record the split point information and record the smallest Gini again.

**continuous\_feature():**  
This function is used to find the smallest Gini among all the properties in the red wine. Use the function continuous\_gini to calculate of all the Gini of all properties and do comparation in order to decide the node in the certain state. gini\_judge is the split point value; result is the value of smallest Gini; col is the column index of this property.

**decisionnode():**

used to create the node on the tree, and the node has 10 parameters:

col\_index is the column index of the node

splitpoint\_value is the value of the spilt point (the point can make the gini index of the certain property minimize

flag is a list which contain the information of the node. The information has the meaning of the node (example: the value is greater or equal to the value of this node), the average value of this property, and the percentage of the data meet the requirement (in other words: the condition is true)

leftchiset is a left subset of the node

rightchiset is a right subset of the node

parent is the connection of this node and its parent node

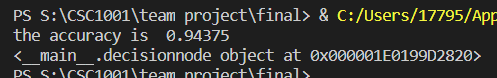
leftchinode is the connection of this node and its left children node

rightchinide is the connection of this node and its right children node

**seed():**this function is used to create the classification tree, the program have already get the root of the tree by using the function of node\_intial. And do precutting at the same time. We can use this function to find the next internal node and create connection between parent node and children node.

**test\_process():**This function is used to read the test file and calculate the correct number of the red wine which quality is greater than 6, and do prediction of all the red wine quality. After that, calculate the number of red wine whose quality are greater than 6 by using the tow list: correct and error. If the prediction is correct, then the number of quality will be appended in correct, else, appended in error. The program can use these two data to calculate the accuracy of this program.

1. **Summary of the results.**

****

1. **Conclusions.**

**When we start to do this project, we learned all the things by ourselves with sources in the reference provided by the instructor, and after that, we found the classification of first dataset is the more difficult, and we want to challenge ourselves, we choose this question. After we are done, the code can’t run normally because of some unknown error. Thus we chose to solve the classification tree of dataset2(the red wine quality) with the initial functions we build for dataset1. (after a few modifications, it works well, and return a high accuracy rate of 94.375%) And at this time, we know how to use CART tree to do classification of data.**

1. **the contribution and workload of each member**

**颜渤东 119010369 the contribution rate 36%**

**Work:**

**1.Initialize data from csv files**

**2.Calculate the gini of continuous type variable with respect to the target variable(like the quality of red wine.**

**3.Adjust and complete the classification tree by recursion based on the the work of 毛邦靖**

**4.Construct the test process and return the accuracy rate with 杨洪萌**

**毛邦靖 118020314 the contribution rate 32%**

**Work:**

**1.Calculate the gini of discrete type variable(when we try to solve problem of Dataset1). The experience of 毛邦靖 contribute a lot in the calculation of gini, helped all the team member learn the difference in gini calculation between continuous type and discrete type**

**2.Build the initial structure of the classification tree**

**3.Try to build a random forest to promote the code and taught other team member the method(by using a dictionary), and the forest contribute a lot in solving the dataset1 problem.**

**杨洪萌 119010378 the contribution rate 32%**

**Work:**

**1.Complete the whole report.**

**2.Help to debug when other two members come across some errors. Find problems efficiently and provide various methods in solving these errors.**

**3.Construct the test process and return the accuracy rate with 颜渤东**